DOCUMENT RESUME

ED 382 669 TM 023 107

AUTHOR Wainer, Howard; And Others
TITLE An Adaptive Algebra Test: A Testlet-Based, Hierarchically-Structured Test with Validity-Based Scoring. Technical Report No. 90-92.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-90-21
PUB DATE 90
NOTE 30p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Algebra; Comparative Analysis; *Computer Assisted Testing; Educational Assessment; Models; Prediction; *Scoring; Secondary Education; Test Construction; Test Format; Test Items; Test Length; *Test Validity
IDENTIFIERS *Hierarchical Models; *Testlets

ABSTRACT

The initial development of a testlet-based algebra test was previously reported (Wainer and Lewis, 1990). This account provides the details of this excursion into the use of hierarchical testlets and validity-based scoring. A pretest of two 15-item hierarchical testlets was carried out in which examinees' performance on a 4-item subset of each testlet was used to predict performance on the entire testlet. Four models for constructing hierarchies were considered. These presentation hierarchies were compared with one another and with an optimally chosen set of four linearly administered items. The comparison was carried out using both the root mean square error and the conditional posterior variance as the criterion. It was found on cross validation that although an adaptive test is everywhere superior to a fixed format test, this superiority is crucially dependent on the quality of the items. When items vary considerably in quality a fixed format test, which uses the best items, can do almost as well as an adaptive test of equal length. Eleven figures and three tables present analysis results. An appendix presents some test items. (Contains 16 references.) (Author/SLD)

RR-90-21

An Adaptive Algebra Test: A Testlet-based, Hierarchically-Structured Test With Validity-based Scoring

Howard Wainer
Charles Lewis
Bruce Kaplan
James Braswell
Educational Testing Service

PROGRAM STATISTICS RESEARCH

TECHNICAL REPORT NO. 90-92

Educational Testing Service
Princeton, New Jersey 08541

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

An Adaptive Algebra Test: A testlet-based, hierarchically-structured test with validity-based scoring[§]

Howard Wainer
Charles Lewis
Bruce Kaplan
and
James Braswell

Educational Testing Service

Abstract

Earlier (Wainer & Lewis, 1990) we reported the initial development of a testlet-based algebra test. In this account we provide the details of this excursion into the use of hierarchical testlets and validity-based scoring. A pretest of two 15 item hierarchical testlets was carried out in which examinees' performance on a four item subset of each testlet was used to predict performance on the entire testlet. Four models for constructing hierarchies were considered. These presentation hierarchies were compared with one another and with an optimally chosen set of four linearly administered items. The comparison was carried out using both the root mean square error and the conditional posterior variance as the criterion. It was found on cross validation that although an adaptive test is everywhere superior to a fixed format test, this superiority is crucially dependent upon the quality of the items. When items vary considerably in quality a fixed format test, which uses the best items, can do almost as well as an adaptive test of equal length.

§ This research was supported by the Educational Testing Service's *Program Research Planning Council*; we are grateful for the help that this provided. We would like to express our thanks to Norma Norris for her help in initial data analysis and to Phyllis Murphy for her aid in gathering the data.

I. Introduction

Modern tests must bear an increasingly heavy burden. No longer is a test a single purpose instrument designed specifically for diagnosis, for placement, or for admission. Now a test score is often used to aid policy makers in decisions about the efficacy of various funding practices. The same score helps to direct students to suitable instruction. It also is used to support admission decisions. Oftentimes the same test is used for all of these purposes even though it is known that a special instrument, specifically designed for that particular purpose, would serve better. Yet, using different measuring instruments for different purposes is not always practical because of the heavy burden that it places on the student. Student time is a scarce and valuable commodity and cannot be used inefficiently. So far, triage decisions about the relative importance of each purpose have had to be made. This has resulted in the use of tests that are suboptimal for some (indeed sometimes all) purposes.

It has been one of the tasks of modern testing to make the testing of individuals more efficient. This is so that the same amount of testing time can be used to accomplish more goals. A pair of such goals might be utilizing the test for the dual purposes of measuring learning and prescribing instruction. One result of this work has been the development of adaptive testing (Wainer et al, 1990), which, when stripped to its essentials, is a system that chooses to administer only those items that will be most informative about an examinee's proficiency, and stops as soon as proficiency has been estimated to within a predetermined level of accuracy. This methodology, in its preliminary implementations, has yielded increases in testing efficiency of about 40% (i.e. the accuracy obtained previously with 100 items could be accomplished with only about 60 items). Of course these are average gains, and vary as a function of the proficiency of the examinee. Examinees whose proficiencies are at the extremes of the distribution typically have greater gains of efficiency; those in the middle, smaller.

Since adaptive testing achieves its increased efficiency through a judicious choice of items, it stands to reason that examinees with different proficiencies will have taken different tests. The glue that has been used traditionally to hold these different tests together (to allow comparisons among individuals who have conceivably all taken different tests) is item response theory (IRT). In IRT, the examinee's observed performance is used to estimate his or her position on an underlying latent variable. Yet this is not the only way that this can be done. Another strategy, devised by Lewis (in preparation), is *validity-based scoring* (VBS) which utilizes methods described by Breiman et al (1984). This methodology uses the reduced length test chosen by an adaptive item choice algorithm as a predictor of performance on some criterion variable. If we wish to think of this in terms of traditional test theory, we could choose the score on the total item pool as the criterion variable. Such an engineering approach has much to recommend it. Aside from not requiring the often untestable assumptions of IRT, it forces the tester to firmly establish what is the validity criterion, and to do appropriate validity studies in advance of scoring any test. In our view, anything that encourages the accomplishment of more validity studies is, *prima facie*, a leg up on any competing method which does not.

The accuracy of any test in a particular proficiency-region is determined by the number (and, to some extent, the quality) of items in that region. In traditional linearly administered tests the items were typically most densely packed where the examinees' proficiency was the densest. This was not true for tests with a known cut-score. In such a situation, test developers try to make the test most accurate in the region of the cut-score. However, for other tests the eventual use was less clear (i.e., college admissions tests have different regions of importance depending upon the specific college; selective schools need more precision at the high end, less selective ones at lower regions). Consequently, general tests were built to provide precision roughly in proportion to the density of the examinee proficiency distribution. What this means practically, is that large-scale tests are better able to discriminate in the middle of the score distribution than they are at the extremes. This has an unfortunate side-effect. Specifically, it means that if the proficiency distribution of a particular subgroup in the examinee population is markedly different from that of the majority, that subgroup will get a less accurate test. In the past, when practical considerations bound us to the use of fixed format tests, we were helpless to correct this problem. Now that is no longer the case.

Adaptive measurement, with appropriately chosen stopping rules, can provide a test that is equally accurate for all examinees. The accuracy of a traditional test was typically measured by its reliability — an aggregate statistic; whereas in an adaptive test some transformation of the information function is quite often used. This is typically shown as a graph plotting the standard error of estimate of proficiency (s.e.[θ]) against proficiency (θ). This is ironic. In a traditional test, the accuracy of the test varied greatly across the range of examinee proficiency, and so such a plot would have been important and useful. In an adaptive test one might use such a plot as a control to be sure that everyone is getting an acceptably accurate test — the test information curve should be relatively flat across the proficiency region in which decisions are to be made. However, once we are assured that it is relatively flat, the aggregate statistics developed for fixed format tests can at last have their restrictive assumptions fulfilled and so would be acceptably accurate. Of course this is exactly the opposite of what is done. In any study that compares the performance of an adaptive test to that of one with a fixed format, the measure of comparison is critical. If one uses an aggregate measure, the efficacy of the adaptive test is denigrated because examinees' proficiencies bunch up in the middle. Thus any test that concentrates its efforts (items) in the middle will do very well. But how does it do in the extremes? An adaptive test will also do well in the middle, but can do equally well in the extremes. However because there are relatively fewer examinees in the tails, its superiority in those regions will be diluted. In this paper we provide some aggregate comparison statistics, but we rely more strongly on a measure of conditional accuracy. For many purposes, this latter measure is the more appropriate. This is discussed more fully in sections III and IV.

II. The Problem

Using computers to administer tests is more expensive initially than paper and pencil methods, but has many advantages (see Wainer et al, 1990, Chapter 1 for a fuller de-

scription). Among these are: ease of modification/updating of item pool, better control of security, and speed of scoring. Dwarfing these, in terms of potential importance, is the capacity to ask qualitatively different kinds of questions that can test qualitatively different traits. Where the multiple-choice format seems ideally suited to a paper and pencil test, a computer administered exam has enormous possibilities that have only just begun to be explored.

If the test is to be given by computer anyway, it seems logical to try and make its administration as efficient as possible. Thus, why not make it adaptive? This minimizes the respondent load or increases the breadth and/or precision of measurement. But item choice algorithms in adaptive testing (Wainer et al, 1990, Chapter 5) tend to focus on the statistical characteristics of the items (the potential gain in Fisherian information) rather than their appropriateness from the point of view of the item's content. This works fine if the unidimensionality assumption explicit in current forms of IRT-based adaptive testing is viable. It is more problematic if this assumption is violated. One solution to this problem is the testlet (proposed by Wainer & Kiely, 1987, and described in greater detail by Wainer & Lewis, 1990). A testlet can be a small number of items that provides a coherent test of a subset of the content domain. The items and their order of presentation can be chosen by a test development expert with the full support of pretest statistics, so that the joint goals of maximal statistical information and content integrity may be served.

The ultimate goal is to prepare a pool of calibrated testlets that can be combined in a variety of ways to efficiently and reliably test a particular domain. Using a testlet rather than an item provides us with a more stable building block for the measurement edifice. How much efficiency is added by making the testlets internally hierarchical? What is the best method for constructing this hierarchy? How stable are the results thus obtained? This study is an attempt to begin to answer these questions.

III. The structure of the study

In this investigation we studied proficiency in algebra. Two 15 item testlets were constructed that spanned *basic algebra skills* (Testlet 1) and *factoring skills* (Testlet 2). The items written for the 15-item testlets were developed to reflect different levels of understanding of a topic (e.g., factoring) and so that various combinations of items could be administered in any sequence without violating the canons of good test construction. We studied both testlets, but since the results were so similar on both of them we report only the results on Testlet 2. The items for Testlet 2 are provided in the appendix. These testlets were given to 2080 ninth and tenth graders. Each student received all 30 items. The resulting data were divided randomly into two sets of 1040 examinees. We shall denote the first set as the *exploratory sample;* the second set was placed in reserve for cross-validation purposes and was called the *confirmatory sample*. The item response data was fit with a three parameter logistic response model using marginal maximum likelihood (Bock & Aitken, 1981; Mislevy & Bock, 1983). With these results in hand each testlet was formed into three separate (but related) hierarchical structures. These structures (and the two others which follow) then formed the basis of a simulated adaptive administration. That is, the complete data were re-analyzed as though students had taken one of these types of tests. In

fact nobody actually responded to the items in any of the ways described here. The structures formed are:

1. **difficulty - structured** — The 15 items of each testlet were ordered by their IRT difficulty parameter (*b*), the first item presented was the middle one (item 8). If this was answered correctly it was followed by item 12, if incorrectly by item 4. The full presentation tree for Testlet 2 is shown in Figure 1. The item numbers in these figures refer to their b-ordering: 1 is the easiest item, 15 the most difficult.

2. **Stepwise optimal tree without replacement** — This tree was formed by choosing, as the start, the item that yielded the minimum posterior variance in the two groups thus formed. The second item on each branch chosen was the one that, when added to the first, minimized the posterior variance in the four groups thus formed. This was continued until a test length of four was reached. The step-wise optimal tree thus devised for Testlet 2 is shown in Figure 2. After an item was used as one node of the tree it was not used again. These trees are formed from the top down and from right to left (i.e., at any level, the rightmost item was chosen first). Consider the third row of items in Figure 2. Item 7 was chosen for examinees who got item 6 wrong because it was the best one remaining since items 2, 4 and 11 were already used Compare with the tree in Figure 3 in which the item choice algorithm could opt for item 4 again.

3. **Stepwise optimal tree with replacement** — The formation rule for this tree is identical to that described in (2), except that after an item was used it was placed back in the pool with the possibility of its being reused on another branch. The tree thus formed for Testlet 2 is shown in Figure 3.

In addition to these three trees two others were formed. The first of these was formed before any data were gathered. We refer to this tree as:

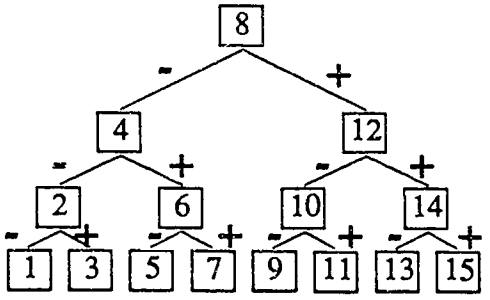
4. **Expert structured** — The 15 items were written and structured by an expert in the development of mathematics tests without any knowledge of the statistical information associated with the items. The presentation tree for Testlet 2 is shown in Figure 4.

Last, but of great practical importance, we must ask, "Do we really need adaptive testing?" How much better do we do with adaptive item selection than we would have done by merely choosing the best four item subset? To examine this possibility, we formed:

5. **Best four item test** — A fixed length test of four items (shown as a tree in Figure 5) by examining all 1365 possible four item tests (15 choose 4 possibilities) and selecting the one that performed best.This procedure yields a result that is a proper subset of optimal trees, and so must be surpassed by method 3, yet using a linearly administered test allows much simpler technology (pencil and paper) than any adaptive test. How

much is our accuracy affected? Is the gain in accuracy associated with adaptive administration worth the additional expense?

Figure 1. Item tree ordered by difficulty



The path to the right indicates a correct response (+): to the left an incorrect one (-).

Figure 2. Item tree sampled without replacement

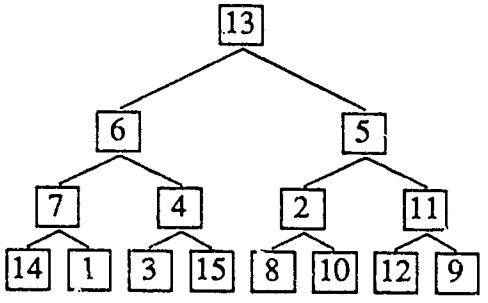


Figure 3. Item tree sampled with replacement

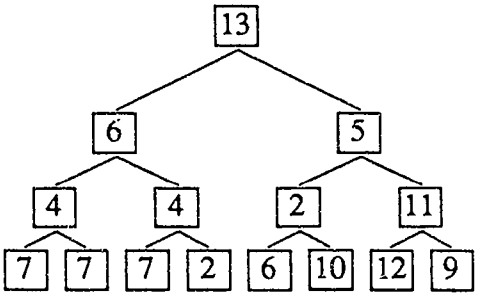


Figure 4. Item tree developed without pretest information

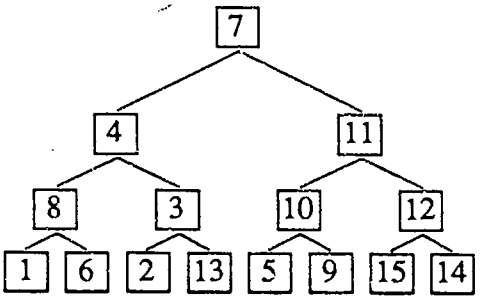
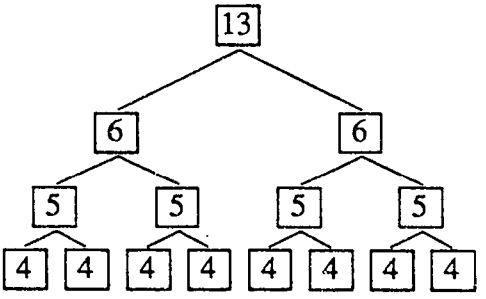


Figure 5. Item tree using only four items



The structure of the criterion variable.

Validity-based scoring can use any suitable criterion variable. For example, using this sort of scheme, one might provide the expected first year college grade point average as an individual's SAT score. In this study we use a criterion score of more modest scope, specifically, we wish to predict the score on the total 15 item testlet from a four-item subtest. We scored the testlets using both the raw score metric (number right) and the latent scale yielded by item response theory. While there are some interesting differences, they are not so profound to justify dual presentation at this time. Thus, we shall present our results within an IRT framework. Specifically, we estimate proficiency (θ) on the entire (15 item) pool and then try to predict it as accurately as possible from the various 4-item branches. The predicted value of proficiency at a particular node in the prediction tree is the mean value of the criterion score for all of those at that node.

IV. The results

All of the results reported in this section are based on what obtained when the trees derived on the exploratory sample were tried on the confirmatory sample. There was some shrinkage in this cross-over, but surprisingly little. We will only discuss the results from Testlet 2.

Shown in Table 1 are the BILOG (Mislevy & Bock, 1983) estimates of the item parameters obtained from the exploratory sample.

Table 1

Item Parameters for the 15 items of Testlet 2

Item Number	*a*	*b*	*c*
1	0.61	-3.28	0.15
2	1.05	-0.82	0.14
3	0.73	-0.76	0.15
4	1.34	-0.64	0.16
5	0.59	-0.41	0.12
6	0.98	-0.35	0.12
7	0.64	-0.26	0.20
8	2.07	0.22	0.08
9	1.86	0.23	0.08
10	1.37	0.74	0.18
11	1.37	0.83	0.10
12	1.11	0.88	0.35
13	1.12	1.27	0.24
14	1.00	1.78	0.08
15	1.01	1.82	0.09

As a first attempt to measure the comparative efficacy of the five item trees described in section III and depicted in Figures 1 through 5 we calculated the η^2 (calculated from the ratio of the between group sums of squares of the groups shown in the various tree diagrams, to the total sum of squares) that obtained with each item presentation tree on the confirmatory sample. These are shown in Table 2.

Table 2.

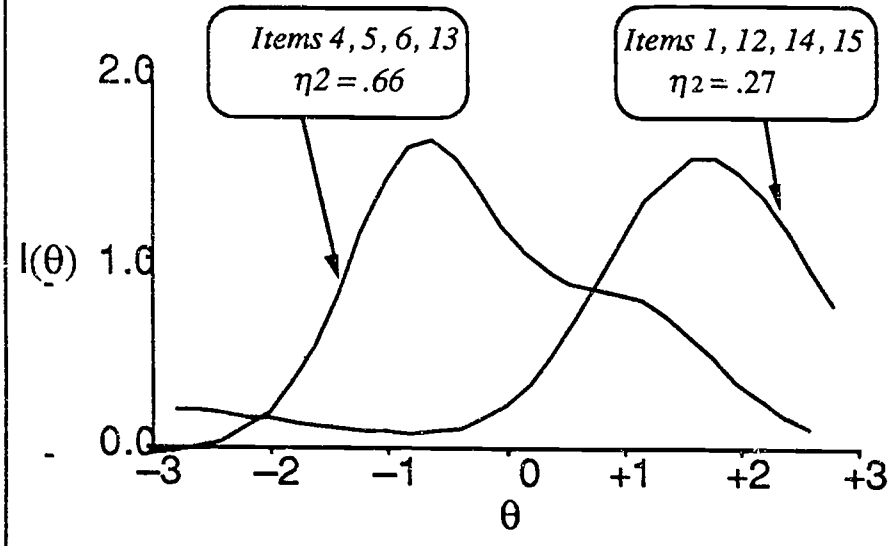
Tree precision - - Represented by η^2

Tree	η^2
1. *difficulty-ordered*	.48
4. *Expert structured*	.54
5. *Best Four*	.66
2. *w/out replacement*	.69
3. *w/ replacement*	.71

These summaries do convey something of the order of finish of the various presentation algorithms, but because they are averaging over the observed proficiency distribution, they do not convey the relative performance of these methods at any particular value of θ. To illustrate this, consider the information functions shown in Figure 6. One represents the best 4-item tree (on the basis of η^2) the other the worst. Note that even though overall we would surely prefer the tree associated with items 4, 5, 6 and 13, higher proficiency examinees would be much better tested with the more extreme items in the worst tree.

Figure 6

The information functions for the best four items and the worst four items on the basis of a multiple squared correlation



Because we are interested in the accuracy of the test over the entire range of proficiencies, we will not use a single summary statistic to characterize a structure's efficacy. Rather, we will follow Birnbaum's (1968) [and more recently, Ramsay's (1982)] advice and use a function. Specifically we will do this in two different ways; this ecumenical approach reflects the two prevailing views of modern statistics. We will examine the theoretically interesting (but unobservable in practice) statistic:

1. *the actual root mean square error* (the square root of the average squared difference between the estimated value of θ and the true value of θ within category) shown as a function of the true θ.

And the more practically useful,

2. *Posterior standard deviation* shown as a function of the estimated posterior mean.

Tree 3 is, in a stepwise sense, the best that we can do. Each person gets the best four item test that can be constructed from the available 15 item pool. How stable its superiority is on cross validation is an empirical question that we examined here; other trees, which are less dependent upon peculiarities in the exploratory (calibration) sample might surpass it on cross-validation. For example, tree 4, which does not depend upon the data

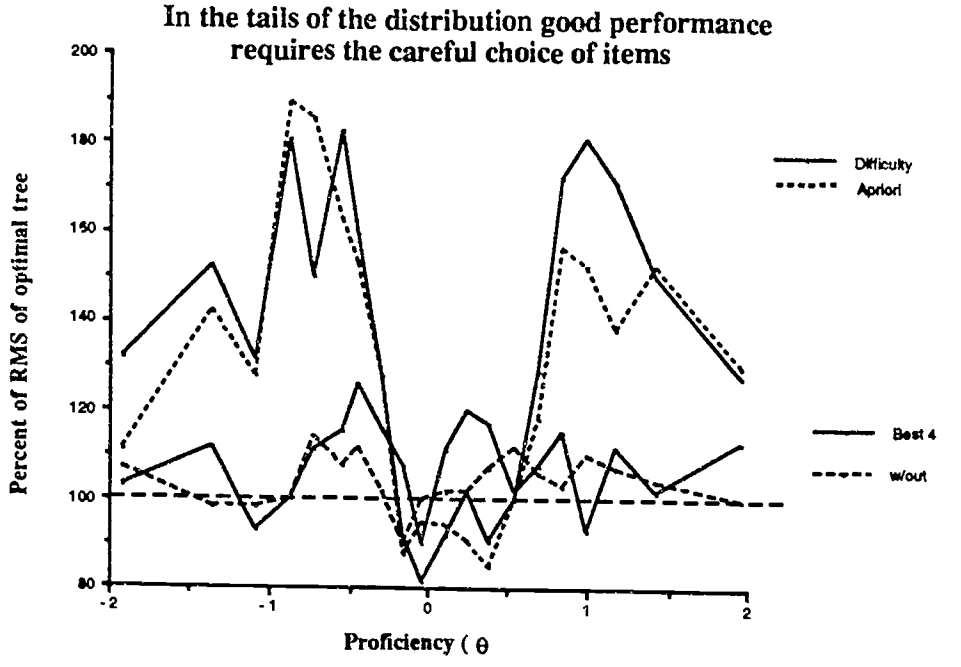
at all for its structure, has no shrinkage on cross validation. Despite these concerns, our finding has been with this test, these examinees, and this sample size, that shrinkage is minimal and the order of finish we saw in the exploratory sample remained the same in the confirmatory sample. In the rest of the discussion we will call tree 3, **the adaptive test.**

Tree 5 is the best fixed format four item test that could be made from this item pool. It does not do as well as tree 3, but it can be administered in a paper and pencil format, and it does do, overall, about two-thirds as well as a test almost four times its length. In the rest of this discussion we will call this the **fixed format test.**

An obvious question is, "Is the marginal gain associated with an adaptive test worth the cost?" Before we address this issue, let us try to understand just how much better the adaptive test is than the fixed format test.

Figure 7 provides a comparison of all five trees, shown as a percentage of their root mean square error of tree 3. Anytime they are greater than 100% it means that tree 3 is superior at that point. This figure shows clearly that the difficulty ordered tree (tree 1) and the tree based on expert judgement, tree 4 (labelled here "a priori") are inferior in the tails. It shows that all trees do about equally well in the middle of the distribution. An examination of these results (and a little *post hoc* intelligence) tells us that there are really only two trees of interest. These are trees 3 and 5.

Figure 7



Shown in Figures 8 and 9 are comparisons of the **fixed format test** (the best four item test) with the optimal **adaptive test.** In Figure 8 is shown the actual root mean square error (in the metric established by the standard deviation of the proficiency distribution) for the two kinds of tests. The increase of error in the tails is exactly what would be expected in any imperfect prediction system. This is caused by the kind of inward regression that has been well known since Galton (1886). While one can tell from this plot that the adaptive test dominates the fixed format test at virtually all levels of proficiency, the extent of this domination is unclear.

Figure 8

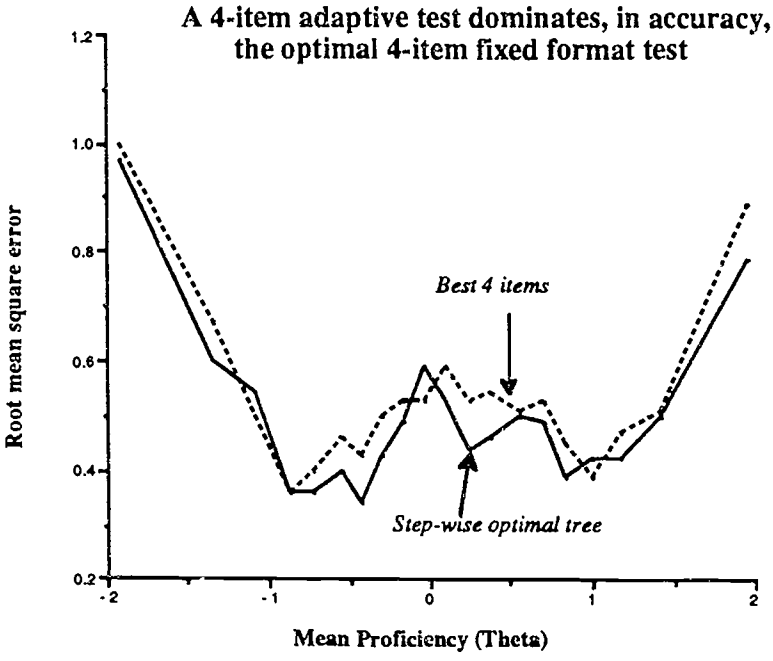
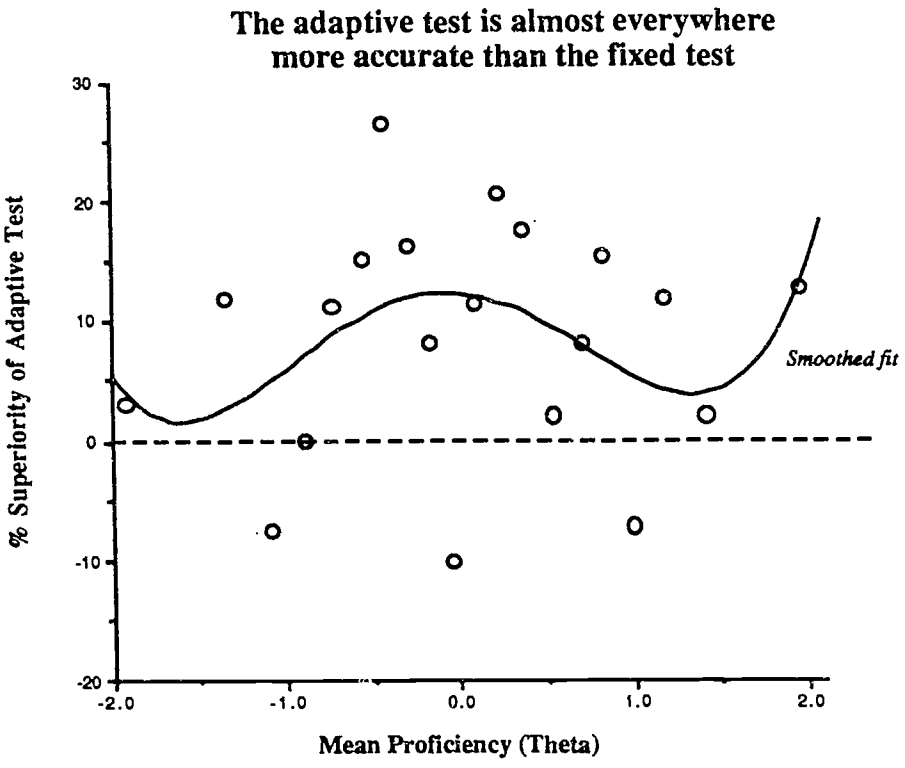


Figure 9 shows the fixed format test's RMS error as a percentage of the adaptive test's error. Thus a point at 10% means that at that point the fixed format test has 10% more error. Similarly a point at -10% indicates that at that point the fixed format test has 10% less error.

Figure 9



Even just a cursory view of Figure 9 indicates that the adaptive test has, on average, about 12% less error than the fixed format test; that sometimes it can be almost 30% more accurate. Somewhat surprising (at least to us) is the shape of the error function. We expected the adaptive test's advantage to be largest in the tails of the distribution. This does not appear to have been the case.

So far we have examined the error of prediction against the true value of proficiency. This is not the situation that we will face in practice. If we knew θ why would we need the test? What the user will know is the mean value of θ that was obtained in the validity study for all individuals at a specific terminal node of the presentation tree. The logical question is "what is the observed variance in θ among individuals with the same performance on the 4-item testlet?" To study this question we can compare the posterior standard deviations for the two kinds of test trees. Such a comparison is shown in Table 3.

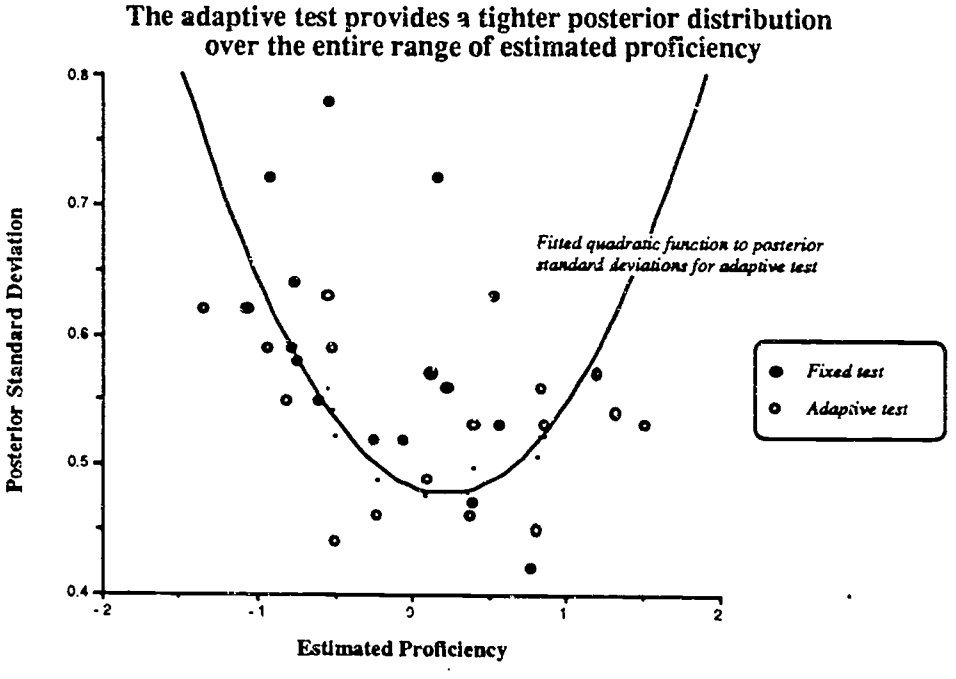
Table 3

Comparing two kinds of testing trees on the location and spread of their posterior distributions

Adaptive tree with replacement			*Best 4-item fixed format test*		
N	*Mean Proficiency*	*Posterior Std. Dev.(θ)*	*N*	*Mean Proficiency*	*Posterior Std. Dev.(θ)*
78	-1.35	0.62	120	-1.06	0.62
57	-0.94	0.59	27	-0.92	0.72
61	-0.81	0.55	25	-0.76	0.64
33	-0.77	0.59	109	-0.74	0.58
29	-0.55	0.63	68	-0.61	0.55
39	-0.52	0.59	21	-0.54	0.78
85	-0.51	0.44	115	-0.25	0.52
149	-0.23	0.46	28	-0.05	0.52
30	0.09	0.49	53	0.12	0.57
43	0.37	0.46	21	0.17	0.72
97	0.40	0.53	43	0.23	0.56
92	0.80	0.45	47	0.39	0.47
133	0.83	0.56	25	0.54	0.63
14	0.86	0.53	90	0.57	0.53
41	1.32	0.54	44	0.77	0.42
59	1.52	0.53	204	1.21	0.57

A careful look at this table shows us that the adaptive test spreads out the examinees a little better (the mean proficiencies of the extreme score groups are more extreme) and has (on the average) about an 8 to 10 percent narrower posterior density. This is a little hard to see within the table. A somewhat clearer look can be had if we graph the estimated posterior means and standard deviations for each of the test formats. In Figure 10 is such a graph which we have augmented by including a fitted curve to the posterior standard deviations of the adaptive test. We can now see that the posterior standard deviations of the fixed test are almost always above this curve; sometimes far above it.

Figure 10


The adaptive test provides a tighter posterior distribution over the entire range of estimated proficiency
Posterior Standard Deviation
0.8
0.7
0.6
0.5
0.4
-2
-1
0
1
2
Estimated Proficiency
Fitted quadratic function to posterior standard deviations for adaptive test
Fixed test
Adaptive test

V. Conclusions

One purpose of this investigation is to illustrate how validity-based scoring can be used with hierarchically structured testlets. We have described this within the context of IRT, but we could have done it in the raw score metric just as easily. IRT is useful in this area, but by no means crucial. The asymptotic results depicted in the information functions were easier to do with IRT, but are subject to all of the well-known caveats associated with the use of any asymptotically-based statistic stemming from a strong measurement model.

Some of the results described here were unexpected. The "almost-no-shrinkage on cross-validation" was a pleasant surprise. We expected the adaptive test derived from the optimal tree structure to prove out to be the best, and it was, but the robustness seen with minor variations in tree structure was a mixed blessing. It was one clue that points us toward the conclusion that the available item pool was neither large enough nor of high enough quality to allow the item choice algorithm the opportunity to really show its stuff. Another finding that hints in this direction was the unexpectedly good performance of the fixed format test.

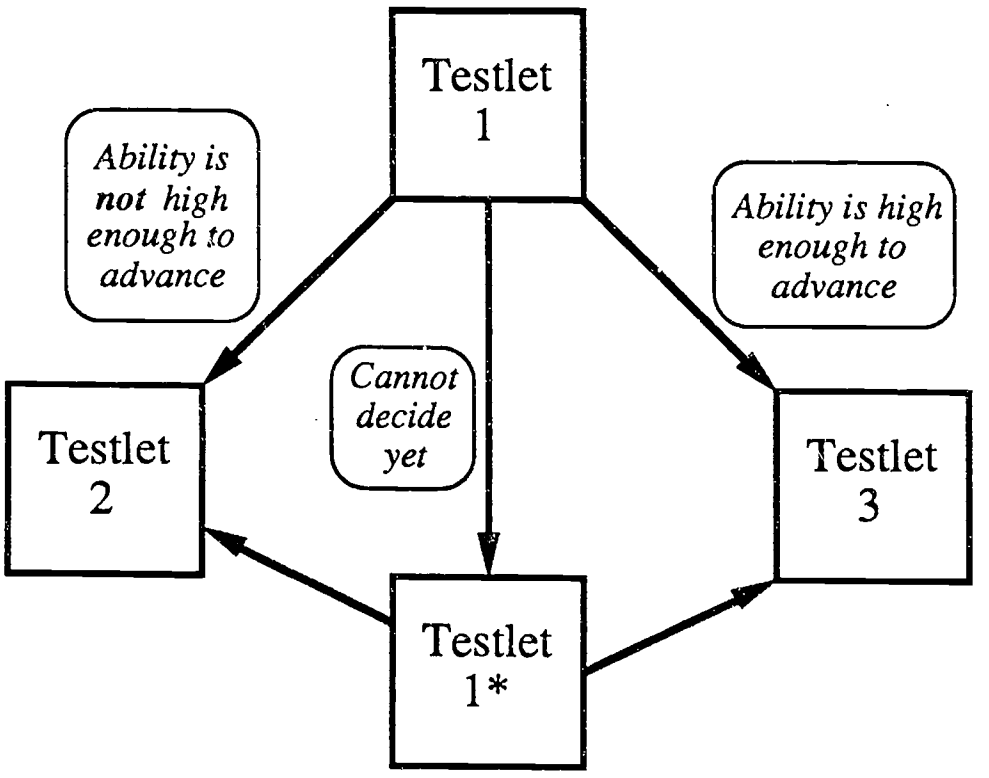
These results support our contention that while adaptive testing is more accurate than fixed format testing, it's hard to justify the expense of computerizing **just** for this increase in accuracy. Rather it should be viewed as an extra benefit that gets thrown in almost for free when computers are used to administer tests. Thus, the logic goes, if we need a computer to administer the test we might just as well make it adaptive; **not** "let's give our test via computer so that we can make it adaptive." Although the modest advantage of adaptive testing within a testlet could be magnified if testlets are strung together hierarchically. We have no data on this as yet.

Nevertheless, enough experience is currently available so that we feel that it would be of some potential use to describe one way that testlets can to be hierarchically combined into a full, adaptive, test. Let us suppose that an examinee takes Testlet 1 (shown in figure 11 below), and for the sake of this example, assume that Testlet 1 was hierarchically structured internally. We are now faced with a decision. Do we administer Testlet 2, which is much easier? Or do we administer Testlet 3, which is much more difficult? In an algebra test, Testlet 1 might be represented by a set of questions on basic algebra skills, Testlet 2 might be simple arithmetic, and Testlet 3 might be the sort of factoring testlet described in this paper. A reasonable concern would be that if we incorrectly consign an examinee to the left branch of this tree, there is no way (short of a re-examination) for that examinee to recover. Indeed, under one conception of the structure we have just described this is true. But it is also true that each allowance for recovery that is made compromises the efficiency of the testing. The challenge is to maintain maximal efficiency while controlling misclassification errors.

A solution that has proved useful in other testing contexts (e.g., Lewis & Sheehan, 1990; Wainer & Lewis, 1990) involves deriving a loss function and from this, setting a cut score. After the completion of Testlet 1 we then decide if the examinee is significantly above or below the cut score. For some examinees this will be decided on the basis of this

testlet; for others it will not. Thus after each testlet we are faced with the trinary decision "Go up," "Go down," or "Keep on testing." This process is shown schematically in Figure 11. Testlet 1* can be thought of as a parallel form of Testlet 1. By adjusting the stringency of the decision process we can control the likelihood of both sorts of errors while still allowing those individuals for whom the decision is clear to progress through the system efficiently. The levels of error that we allow ourselves will determine the number of parallel forms of each testlet that we need to construct. Of course the posterior proficiency distribution that emerges from each testlet becomes the prior for the next. This allows the testing to progress with increasing speed.

Figure 11



VI. References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability (Pp.397-479). In Lord, F. M., & Novick, M. *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46*, 443-449.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431-444.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute, 15,* 246-263.

Lewis, C. (in preparation). *Validity-Based Scoring*. Princeton, N.J.: Educational Testing Service.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14,* xxx-xxx.

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison Wesley.

Mislevy, R. J., & Bock, R. D. (1983). *BILOG: Item and test scoring with binary logistic models* [computer program]. Mooresville, IN: Scientific Software.

Monk, J. J., & Stallings, W. M. (1970). Effect of item order on test scores. *Journal of Educational Research, 63*, 463-465.

Ramsay, J. O. (1982). When data are functions. *Psychometrika, 47,* 379-396.

Rock, D. A. (April, 1988). *Multidimensionality as knowledge hierarchies*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana.

Rosenbaum, P. R. (1988). A note on item bundles. *Psychometrika, 53*, 349 -360.

Wainer, H., Dorans, N., Flaugher, R., Green, B., Mislevy, R., Steinberg, L., & Thissen, D. (1990). *Computerized Adaptive Testing: A Primer*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*, xxx-xxx.

Appendix

The Items for Testlet 2

1. Which is **NOT** a factor of 36?

(A) 2
(B) 4
(C) 6
(D) 8
(E) 12

2. $(x - 3)^2 =$

(A) $x^2 - 9$
(B) $x^2 - 6x + 9$
(C) $(x - 3)(x + 3)$
(D) $x^2 - (6x)^2 - 3^2$
(E) $x^2 - 6$

3. $x^2 + 5x + 6 =$

(A) $(x - 3)(x - 2)$
(B) $(x + 6)(x - 1)$
(C) $(x + 2)(x + 3)$
(D) $(x + 6)(x + 1)$
(E) $(x + 5)(x + 1)$

4. Which expression is **equal to** $x(6x + 4) + 2x - 4$?

(A) $7x^2 + 2x$
(B) $10x^2 + 2x - 4$
(C) $6x^2 + 2x$
(D) $8x^2$
(E) $6x^2 + 6x - 4$

5. If $3x^2 - 12x = 0$, $x =$

(A) -4 only
(B) 0 only
(C) 4 only
(D) 0 or -4
(E) 0 or 4

6. The **greatest common factor** of $34x^3y + 51x^2y^2$ is

(A) $17\,x^2\,y$
(B) $3\,x\,y$
(C) $x\,y$
(D) $17\,x\,y$
(E) $17\,x^3\,y^2$

7. What is the **missing factor**?
$6x^2 - x - 35 = (3x + 7)\,(\ ?\ )$

(A) $(3x + 5)$
(B) $(3x - 5)$
(C) $(2x + 5)$
(D) $(2x - 5)$
(E) $(2x - 7)$

8. The **missing factor** in $21a^6b^3 = 7a^2b\,(\ ?\ )$ is

(A) $3\,a^4$
(B) $3\,(a^4 + b^2)$
(C) $3\,a^3\,b^2$
(D) $3\,a^4\,b^2$
(E) $14\,a^3\,b^2$

9. If $x^2 - 14x + k^2 = (x - k)^2$, which of the following is the **value of** k ?

(A) 2
(B) 7
(C) 14
(D) 28
(E) 49

10. One **factor** of $2x^2 - 11x - 21$ is

(A) $(2x - 7)$
(B) $(2x - 3)$
(C) $(2x + 3)$
(D) $(x - 3)$
(E) $(x + 3)$

11. One **factor** of $32x^3 - 8x$ is

(A) $(4x^2 + 1)$
(B) $(2x + 1)$
(C) $(4x - 1)$
(D) $(2x - 2)$
(E) $(x - 2)$

12. If $(4x - 1)^2 - (2x + 3)^2 = (6x + 2)(2x - 1)$ **then** $x =$

(A) 1/4
(B) 2
(C) -3/2
(D) -1/2
(E) -1/3

13. What is the **missing term**?

$(8n + 5)^2 = 64n^2 + \underline{?} + 25$

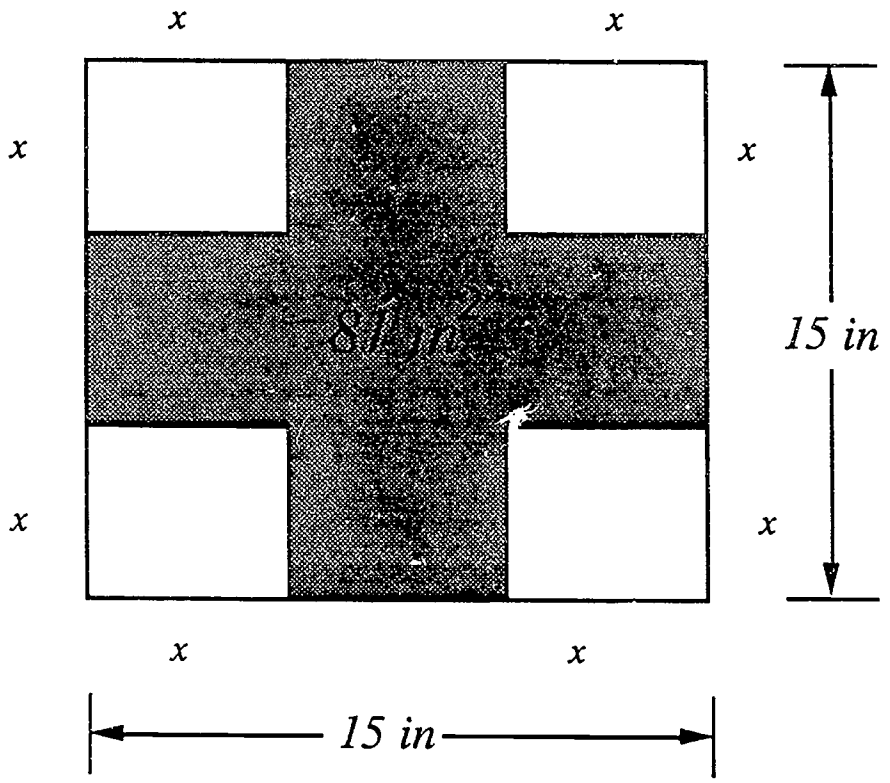
(A) 0

(B) 13 n

(C) 20 n

(D) 40 n

(E) 80 n



Note: Figure not drawn to scale

14. The square piece of cardboard depicted above is 15 inches on a side. Squares of sides x inches are cut out of the four corners, leaving an area of 81 square inches of cardboard. **What is the value of x?**

(A) 2
(B) 3
(C) 6
(D) 9
(E) 12

15. The product of two consecutive, positive odd numbers is 63.

What is the sum of these two numbers?

(A) 9

(B) 16

(C) 17

(D) 18

(E) 19